T-REX: A Framework to Build Trustworthy Recommenders of Evidence Explanation

Andrea Fedele[1,2,†], Mattia Franchi de' Cavalieri[3,4,†], Cristiano Landi[1,2], Clara Punzi[5,1,2,*,†] and Stefano Tramacere[7,1,†]

[1]University of Pisa, Pisa, Italy
[2]KDD Lab, ISTI-CNR, Pisa, Italy
[3]The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
[4]Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy
[5]Scuola Normale Superiore, Pisa, Italy
[6]Scuola Superiore Sant'Anna, Pisa, Italy
[7]LIDER-Lab, Scuola Superiore Sant'Anna, Pisa, Italy

Abstract

The initial enthusiasm for eXplainable Artificial Intelligence (XAI) has been tempered by concerns about the effectiveness and reliability of its explanations. Studies show that some explanations are no more reliable than random ones. Tim Miller suggests a paradigm shift in XAI to address issues of cognitive biases, such as automation bias, which can affect decision-making processes. He advocates for hypothesis-driven support systems to align AI explanations with human cognitive processes. Addressing these issues, we propose the Trustworthy Recommenders of Evidence eXplanations (T-REX) framework. This approach aims to enhance XAI by moving from statistical explanations to those based on trustworthy scientific evidence, enabling AI systems to tackle complex tasks more effectively.

Keywords

Human-Machine Interaction, Explainable AI, Trustworthy AI

1. Introduction

Following the initial boom in eXplainable Artificial Intelligence (XAI), the scientific community began questioning the effectiveness, reliability, and social impact of such explanations. In [1], the authors experimentally demonstrate that the faithfulness and stability of some explanations can be comparable to or even worse than random explanations. Additionally, in [2, 3], Miller advocates for a paradigm shift in XAI to address the concerns about the reliability of automated systems. These systems can be compromised by cognitive biases, such as over-reliance, where users place excessive trust in system recommendations, or under-reliance, where users distrust the system's outputs. Additionally, issues with reliability may arise due to misalignment between the AI system explanations and the cognitive processes used by humans in decision-making, which Miller suggests handling by developing hypothesis-driven support systems [2].

Critical domains require human-AI collaboration, where appropriate reliance is key to the successful use of the technology [4]. In recent years, we witnessed the rise of Large Language Models (LLMs), which are rapidly revolutionizing our society by enabling new types of human-machine interactions [5]. While LLMs are at the forefront of research, they can produce hallucinations (i.e., incorrect outputs), especially when queried about information that are not included in the training set [6]. In such cases, it is critical that humans interacting with LLMs

Discovery Science - Late Breaking Contributions 2024
*Corresponding author.
†These authors contributed equally.
clara.punzi@sns.it (C. Punzi)
0009-0007-0467-0967 (A. Fedele); 0000-0003-0893-1048 (M. F. d. Cavalieri); 0000-0003-4907-9728 (C. Landi); 0000-0002-1366-9833 (C. Punzi); 0009-0004-9801-8942 (S. Tramacere)

understand their limitations in order not to fall into over-reliance in the (not-so-rare) case of incorrect information.

In order to foster such synergistic human-machine collaboration, we advocate for an evidence-driven XAI methodology, which builds upon the strengths of explainability techniques and, at the same time, mitigates its potential lack of persuasiveness or informative content. Leveraging the authors' diverse multidisciplinary expertise (i.e., computer science, biomedical engineering, and law), we propose a framework to build Trustworthy Recommenders of Evidence eXplanations (T-REX), advancing the XAI field from statistical-based explanations to trustworthy community-approved scientific evidence-based explanations. For this purpose, we suggest exploiting reliable and traceable sources, such as scientific literature or World Health Organization (WHO) publications, as a privileged knowledge base for the system, which additionally provides a valuable layer of transparency and accountability, especially in high-risk applications. Additionally, implementing explainability measures will not only improve the appropriate utilization of the AI system by human actors[1] to support their decision-making but also boost the transparency and human oversight of the entire AI system as required by the AI Act (EU Reg. 1689/2024), specifically in Art. 13 (1) and (3)(b)(iv), and Art. 14 (4)(c).

2. T-REX Framework

T-REX is a human-AI hybrid decision-making system [7] where the human actor synergistically interacts with the machine; the framework objective is to support the human-AI decision process by means of community-approved evidences, like scientific literature or WHO guidelines. Figure 1 provides a graphical illustration of the framework in a medical use case scenario. Specifically, T-REX aims at moving from an aseptic <ML outcome, Explanation> bundle towards an approach that cross-validates and enriches such pair with reputable sources authored by domain experts, where the human-in-the-loop interaction is facilitated in the exploration of various hypotheses. In order to satisfy the needs of the human actors and help them analyze doubts and possibilities, such hypotheses may be validated through multiple interactions with the machine. To further boost a critical evaluation of the hypotheses, T-REX is designed to provide humans not only with evidence in support of them but also with those that contradict them (i.e., Supportive and Contrastive Evidence).

T-REX framework involves 4 main components: (i) a human actor, (ii) a ML classifier f, (iii) an explainability technique XAI, and (iv) an evidence retrieval and evidence classifier g. Specifically, for a given query input x, f outputs a set of predictions $\hat{Y}$ along with their confidence C score (e.g., the predicted class probability), formally $(\hat{Y}, C) = f(x)$. After that, the XAI explainer returns an explanation $e_i = XAI(x, \hat{y}_i)$ where y_i is the prediction selected by the human actor, which could be based on its hypotheses only or according to the model confidence.Finally, a composite function g employs the explanation e_i to construct a query and retrieve relevant evidence from reputable sources. It then classifies the selected documents as supportive S or contrasting C concerning the hypothesis under analysis $\hat{y}_i$. This process and the retrieved evidence support the human actor in making the final decision.

Medical Use Case Scenario. The T-REX framework has the potential to make a significant impact in the medical field by offering an innovative system for evidence-based clinical decision-making in diagnostics. The process starts when a human actor inputs patient data x for analysis by the machine. The first algorithmic component f performs a standard classification task, estimating the probability distribution over a set of possible outcomes. Although T-REX is model-agnostic with respect to the choice of f, in this use case, we choose the model proposed in [8], where the authors developed a model for detecting chronic diseases. They trained an

[1]Defined as deployers in the AI Act.

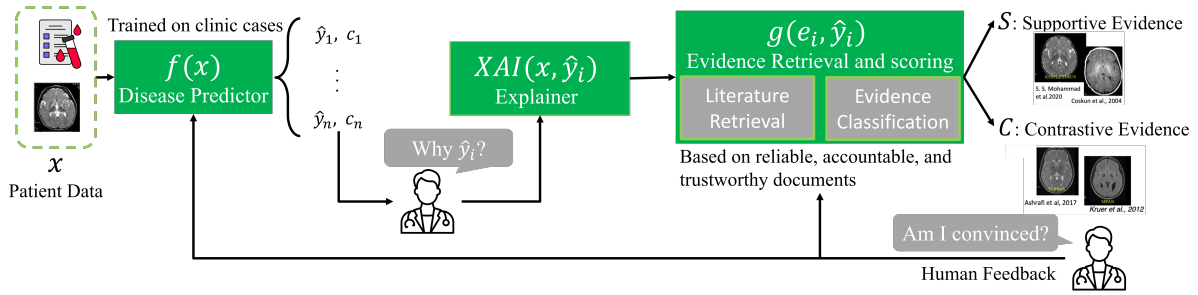

Figure 1: Overview of the proposed framework

artificial Neural Network (NN) classifier on different tabular datasets for each condition: breast cancer, diabetes, heart attack, hepatitis, and kidney disease.

Let us suppose the doctor chooses to investigate a specific prediction $\hat{y}_i$. T-REX generates a statistical-based explanation e_i for $\hat{y}_i$ by supplying the importance scores of the features underlying the predictor f results. This explanation is generated by the XAI component, which could be implemented, for instance, using the well-established SHapley Additive exPlanations (SHAP) technique [9, 10]. Remarkably, different choices of explanation method could be taken depending on the dataset and task of the specific use case scenario.

The final component of the framework, represented as the function g, is composed of two modules. The first module exploits the statistical-based explanation e_i to query the knowledge base and retrieve the relevant documents with respect to the prediction $\hat{y}_i$ chosen by the doctor; such retrieval is part of a well-known task in the computer science literature, known as semantic search [11, 12, 13, 14]. The knowledge base should be composed of reliable, accountable, and trustworthy documents, such as publications by WHO and scientific literature. This is a very crucial part of the proposed framework: having the possibility of relying on such documents enables the doctor to switch from data to medical science, i.e., understand the machine's decision on the basis of medical evidence and not on the basis of some statistical distribution in a dataset which could contain errors. The second component of g uses $\hat{y}_i$ to group the retrieved documents as either supporting or contrasting the hypothesis under analysis, specifically the detection of the disease $\hat{y}_i$. A potential solution involves leveraging a state-of-the-art NLP techniques, such as sentiment analysis, where the goal is to determine whether the sentiment expressed in the text indicates a favorable or unfavorable stance toward a specific subject.

Overall, the T-REX framework encompasses multiple stages of human-AI interactions. First of all, the human actors can choose which potential disease prediction to investigate to test their hypothesis. Furthermore, they can explore different prediction paths and trigger a human-feedback loop by changing the hypothesis under analysis or by tweaking the query derived from the explanation to incorporate additional knowledge.

3. Conclusion

This paper introduces T-REX, a framework that facilitates a more transparent and trustworthy decision-making process in critical applications, such as healthcare, where the cost of errors is critical. Combining statistical-based traditional explanations with evidence retrieved from trusted sources enables human decision-makers to critically evaluate both supportive and contrasting evidence related to AI predictions, potentially mitigating the risk of cognitive biases. Additionally, AI systems using the T-REX framework should facilitate compliance with the legal requirements regarding transparency outlined in the AI Act. The overall interface of the system will benefit from further investigation; currently, we are considering using a traditional web-like interface, as in [15, 16], to avoid the risk of hallucination that could occur with modern Retrieval-Augmented Generation LLMs interfaces [5, 13]. We believe that the combination of explainability and

evidence-based reasoning offered by T-REX represents a promising direction for creating more reliable, trustworthy, and accountable AI systems in the future.

Acknowledgments

This work is partially supported by the European Union NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence) FAIR (Future Artificial Intelligence Research), and "SoBigData.it" - Prot. IR0000013, Res. Infr. G.A. 871042 SoBigData++, G.A. 761758 Humane AI, G.A. 952215 TAILOR, ERC-2018-ADG G.A. 834756 XAI.

References

[1] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, Openxai: Towards a transparent evaluation of model explanations, in: NeurIPS, 2022.

[2] T. Miller, Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI, in: FAccT, ACM, 2023, pp. 333–342.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[4] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors: The Journal of the Human Factors and Ergonomics Society 46 (2004) 50–80. URL: https://doi.org/10.1518/hfes.46.1.50_30392.

[5] Y. Huang, J. Huang, A survey on retrieval-augmented text generation for large language models, CoRR abs/2404.10981 (2024).

[6] J. Yao, K. Ning, Z. Liu, M. Ning, L. Yuan, LLM lies: Hallucinations are not bugs, but features as adversarial examples, CoRR abs/2310.01469 (2023).

[7] C. Punzi, R. Pellungrini, M. Setzu, F. Giannotti, D. Pedreschi, Ai, meet human: Learning paradigms for hybrid decision making systems, CoRR abs/2402.06287 (2024).

[8] J. Rashid, S. Batool, J. Kim, M. Wasif Nisar, A. Hussain, S. Juneja, R. Kushwaha, An augmented artificial intelligence approach for chronic diseases prediction, Frontiers in Public Health 10 (2022) 860396.

[9] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: NIPS, 2017, pp. 4765–4774.

[10] J. Allgaier, L. Mulansky, R. L. Draelos, R. Pryss, How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare, Artificial Intelligence in Medicine 143 (2023) 102616.

[11] H. Bast, B. Buchhold, E. Haussmann, et al., Semantic search on text and knowledge bases, Foundations and Trends® in Information Retrieval 10 (2016) 119–271.

[12] R. Bordawekar, O. Shmueli, Using word embedding to enable semantic queries in relational databases, in: Proceedings of the 1st workshop on data management for end-to-end machine learning, 2017, pp. 1–4.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[14] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). arXiv:2401.08281.

[15] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, in: FAT*, ACM, 2020, pp. 629–639.

[16] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling, in: ISCC, IEEE, 2021.